# Master thesis

## Spatio-temporal Gaze Estimation for Human-Machine and Human-Human Interaction

Alberto López Sánchez
Director: Sergio Escalera
Codirector: Cristina Palmero

Master in artificial intelligence
Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat Rovira i Virgili

January 31, 2020

# Outline

# Contextualization

- **Gaze** is an essential component in non-verbal communication

# Contextualization

- **Gaze** is an essential component in non-verbal communication
- Estimate gaze using **conventional** hardware as RGB camera is very interesting to analyze Human-Human and Human-Computer interactions.

# Contextualization

- **Gaze** is an essential component in non-verbal communication
- Estimate gaze using **conventional** hardware as RGB camera is very interesting to analyze Human-Human and Human-Computer interactions.

## Objective

Exploit the **spatio-temporal** nature of eye dynamics to improve the regression of gaze vectors in **mid-distance** scenario

# Gaze estimation methods

- Model-based
  - Require **high-resolution** images or calibration **specific** parameters to estimate personal eye parameters.

# Gaze estimation methods

- Model-based
  - Require **high-resolution** images or calibration **specific** parameters to estimate personal eye parameters.
- Appearance-based
  - Can be used in **low-resolution** images or a **mid-distance** scenario.
  - Learn a direct mapping from intensity images or extracted eye features to gaze directions.

# State-of-the-art gaze estimation techniques

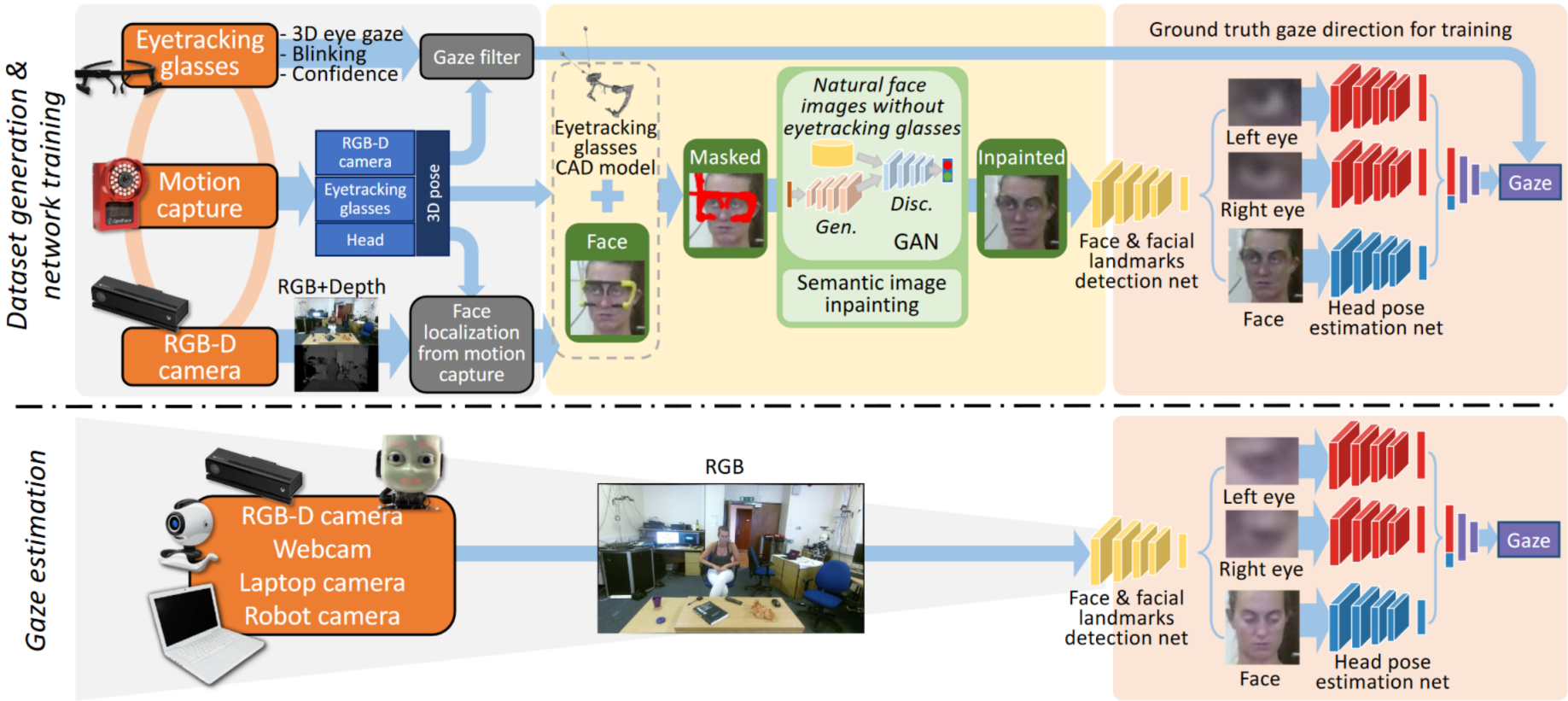



Figure: Work of Fischer *et al.* RT-GENE (Real-Time Eye Gaze Estimation in Natural Environments) Architecture overview.

# State-of-the-art spatio-temporal methods (I)

Work of **Wang** *et al.* Neuro-inspired Eye Tracking with Eye Movement Dynamics.

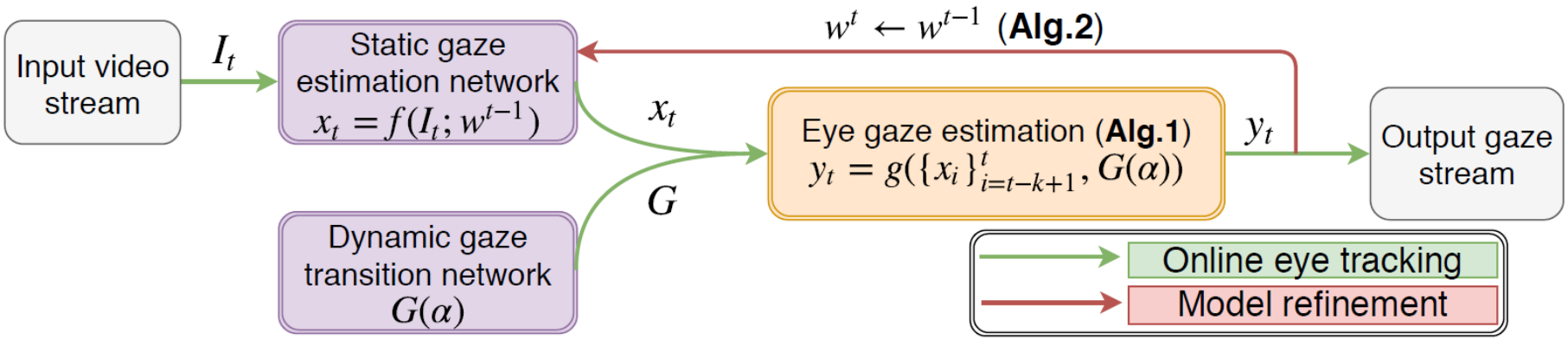



Figure: Overview of the proposed system. Combine static gaze estimation network with dynamic gaze transition network to obtain better gaze estimation.

# State-of-the-art spatio-temporal methods (II)

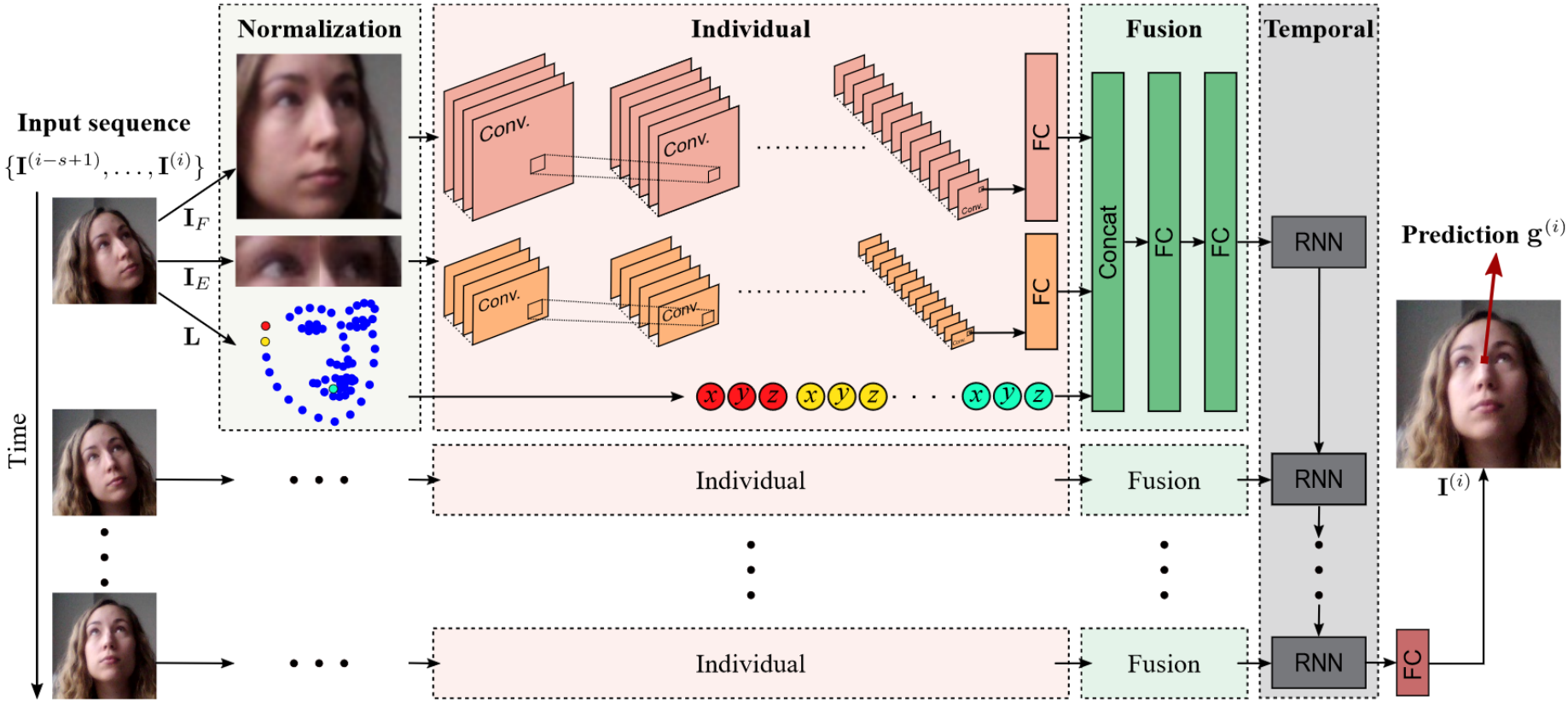



Figure: Palmero *et al.* work. A multi-stream CNN jointly models full-face, eye region appearance and face landmarks from still images. The combined extracted features from each frame are fed into a recurrent module to predict last frame's gaze direction.

# Contribution

- We want to improve the state-of-the-art taking the advantages of temporal information. In particular, using **optical flow**.

# Contribution

- We want to improve the state-of-the-art taking the advantages of temporal information. In particular, using **optical flow**.
  - It defines the **local movement** of consecutive images regions.

# Contribution

- We want to improve the state-of-the-art taking the advantages of temporal information. In particular, using **optical flow**.
  - It defines the **local movement** of consecutive images regions.
  - Defines a **compact representation** of motion.

# Contribution

- We want to improve the state-of-the-art taking the advantages of temporal information. In particular, using **optical flow**.
  - It defines the **local movement** of consecutive images regions.
  - Defines a **compact representation** of motion.
  - **2D kernels** can directly learn motion information.

# Contribution

- We want to improve the state-of-the-art taking the advantages of temporal information. In particular, using **optical flow**.
  - It defines the **local movement** of consecutive images regions.
  - Defines a **compact representation** of motion.
  - **2D kernels** can directly learn motion information.
  - Has less **parameters** and less prone to **overfitting** than other deep learning techniques.

# Contribution

- We want to improve the state-of-the-art taking the advantages of temporal information. In particular, using **optical flow**.
  - It defines the **local movement** of consecutive images regions.
  - Defines a **compact representation** of motion.
  - **2D kernels** can directly learn motion information.
  - Has less **parameters** and less prone to **overfitting** than other deep learning techniques.

# Outline

# EYEDIAP Dataset

We use this dataset because the data is the more realistic one for the task we want to solve.

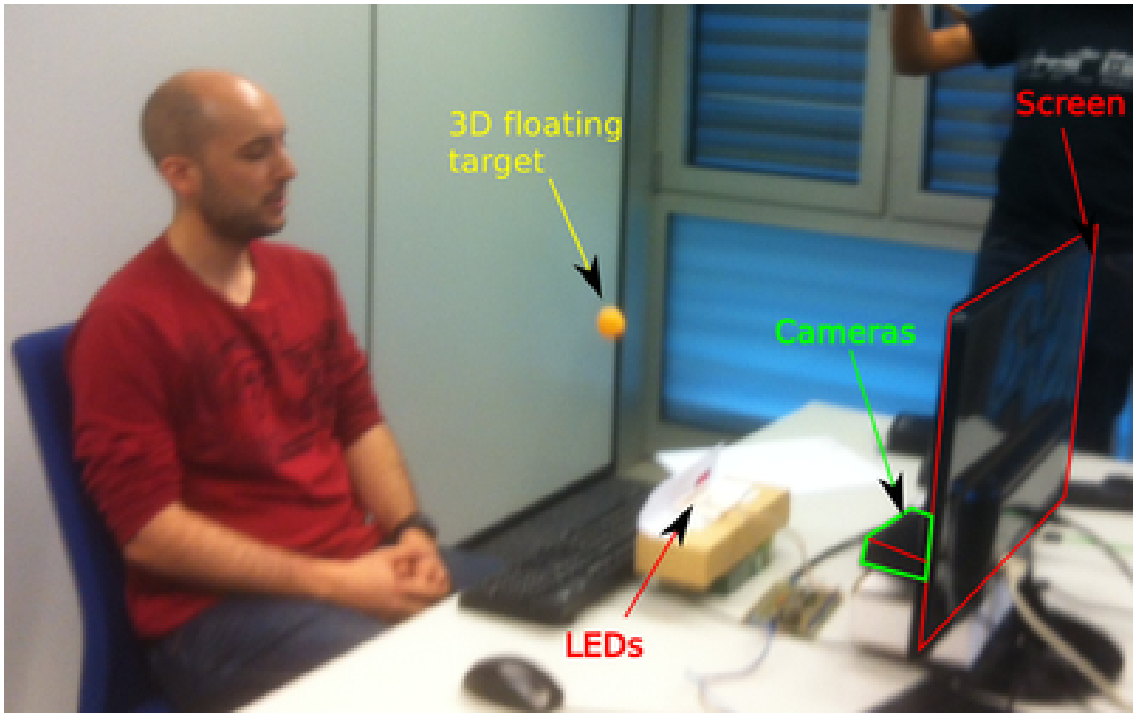



Figure: Setup of the recording room for the EYEDIAP dataset where can be seen the subject, the floating target, the screen target and the cameras. (from [2])

# EYEDIAP examples

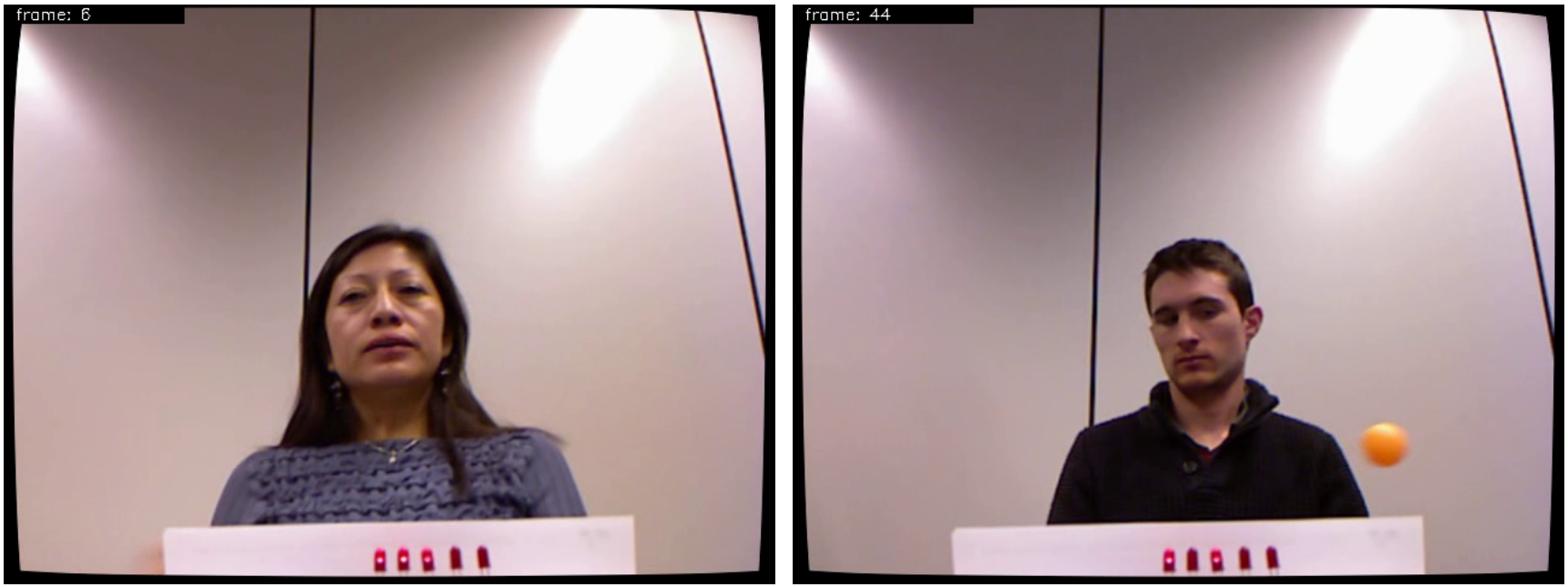



Figure: Two sample images of subjects of the EYEDIAP dataset. (from [2])

# Outline

# Features

- Face and eyes
- Optical flow of the face and eyes
- Face landmarks

# Face and eyes

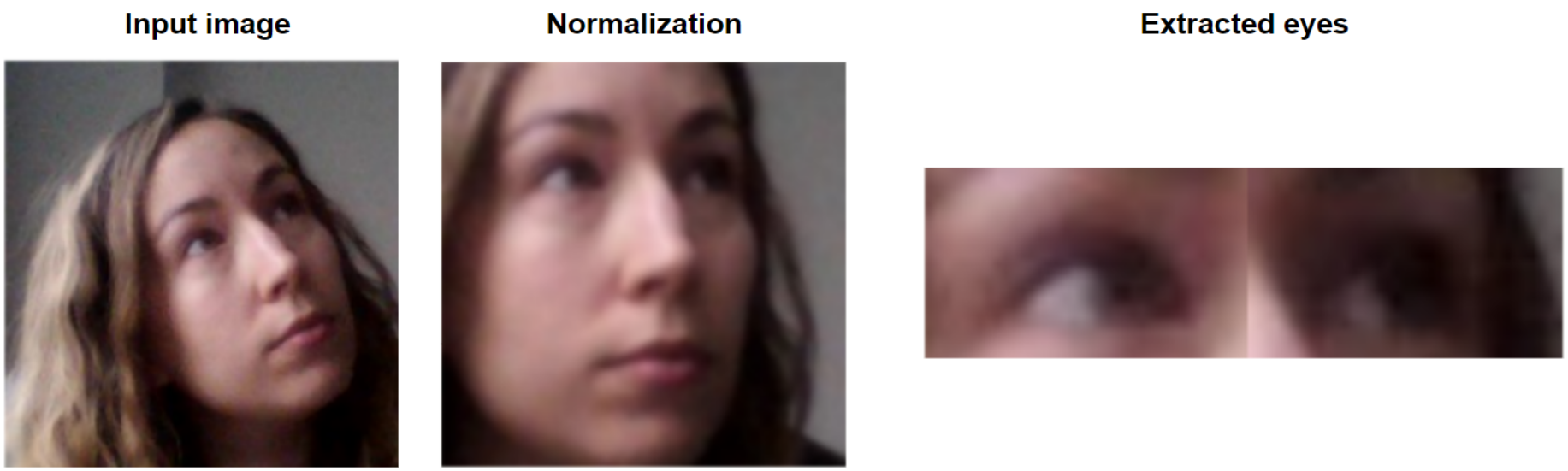
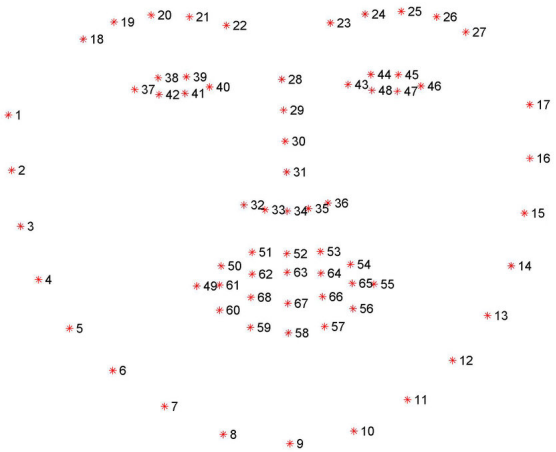



Figure: An example of the original face image, the normalized face and the eyes region extracted. Face images extracted from [4].

# Landmarks

Landmarks are extracted using the state-of-the-art method of Bulat and Tzimiropoulos [3]



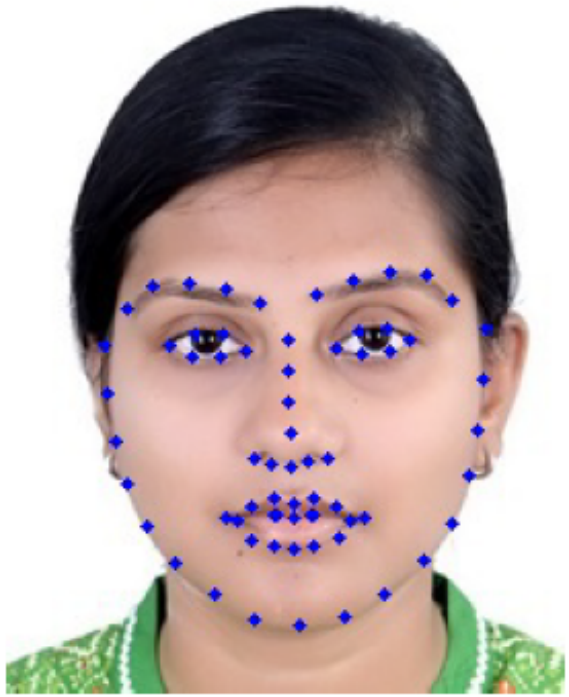



Figure: Visualization of the coverage of the 68 face landmarks.

# Optical flow example in EYEDIAP

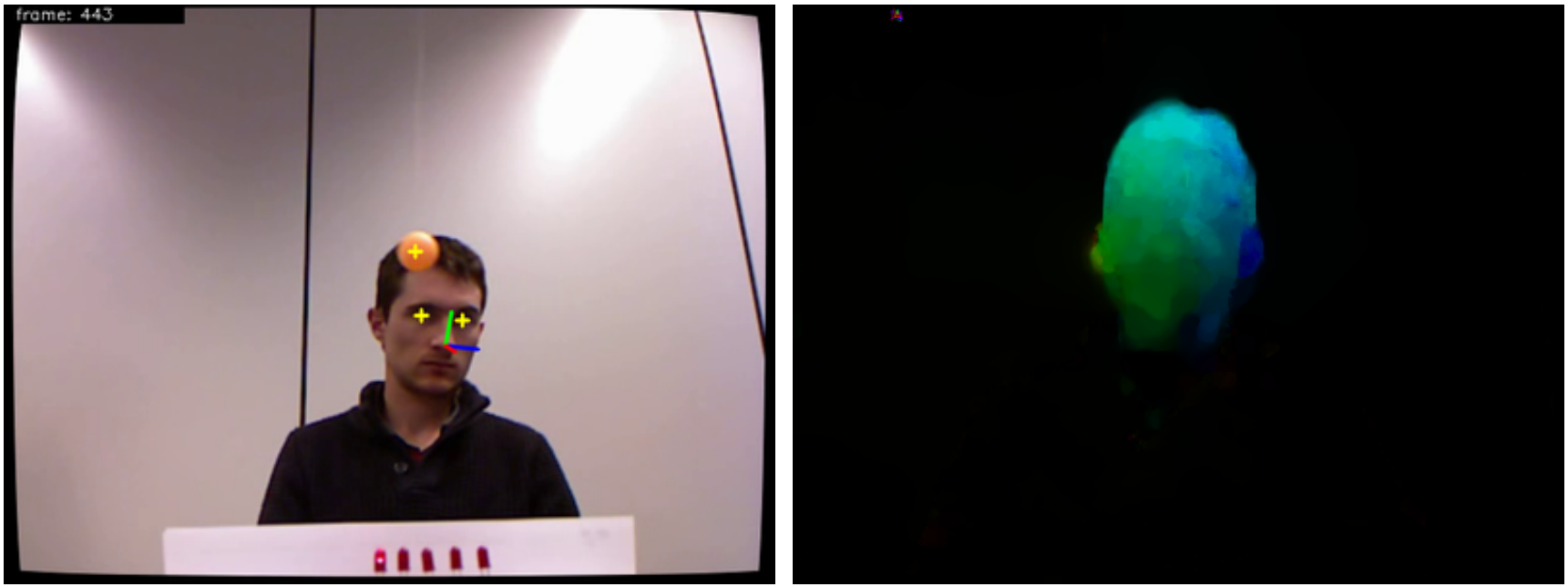



Figure: Example of visualization of optical flow between two sequential frames of the dataset

# Outline

# Static models

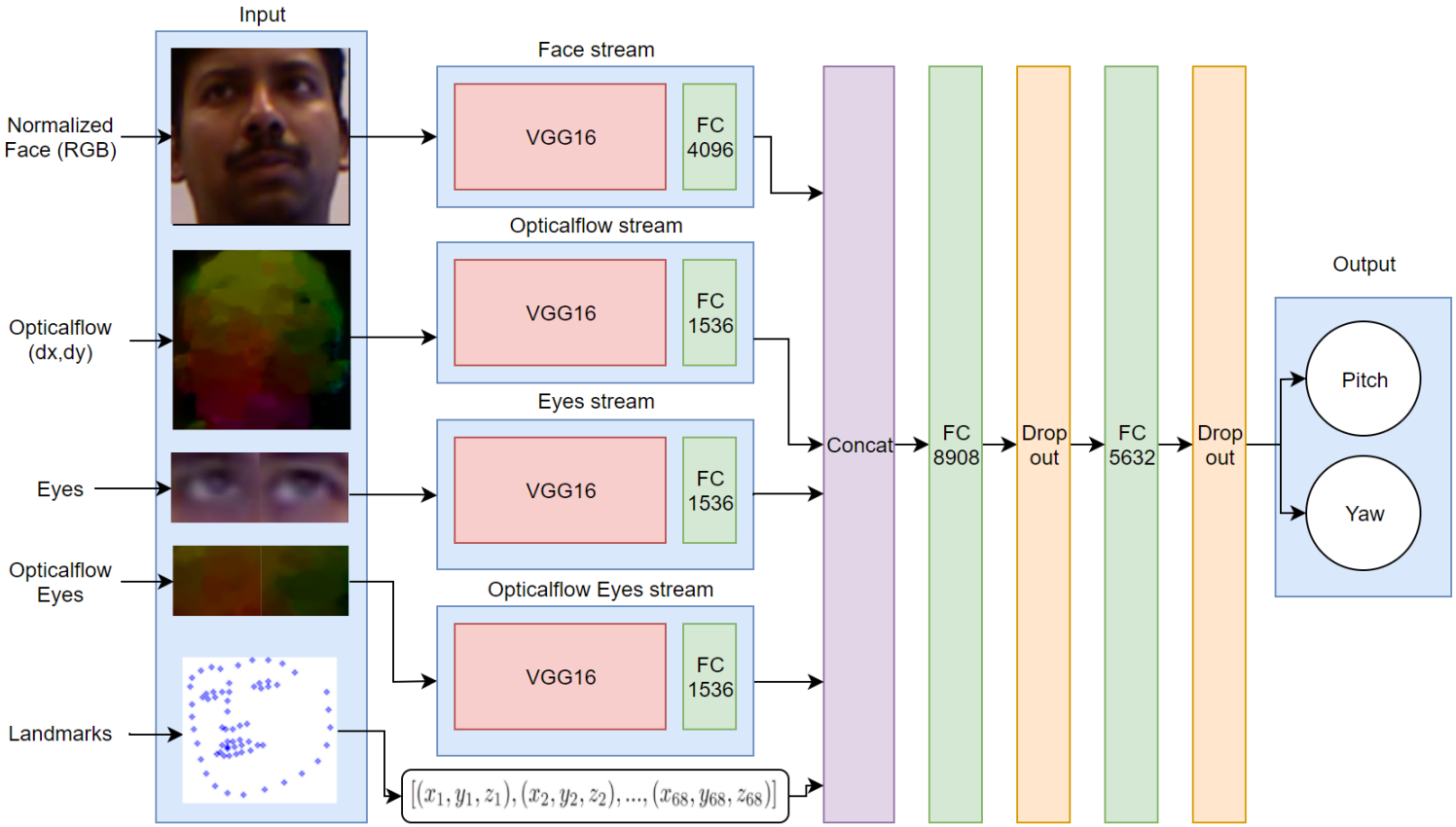



Figure: Architecture of the four stream network model for face, eyes, optical flow of face, optical flow of eyes and face landmarks.

# One Stream

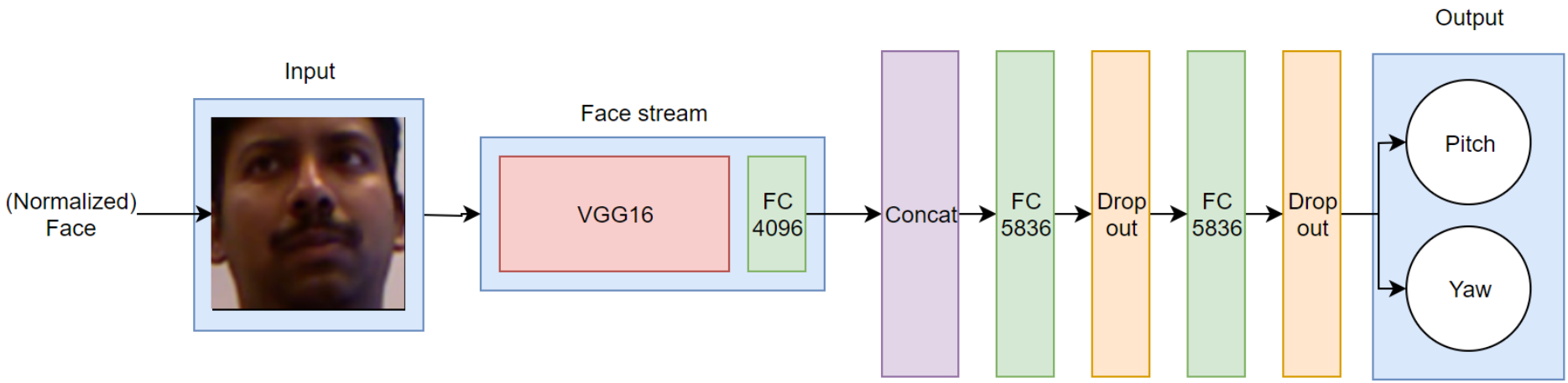



Figure: One stream model.

# Two streams

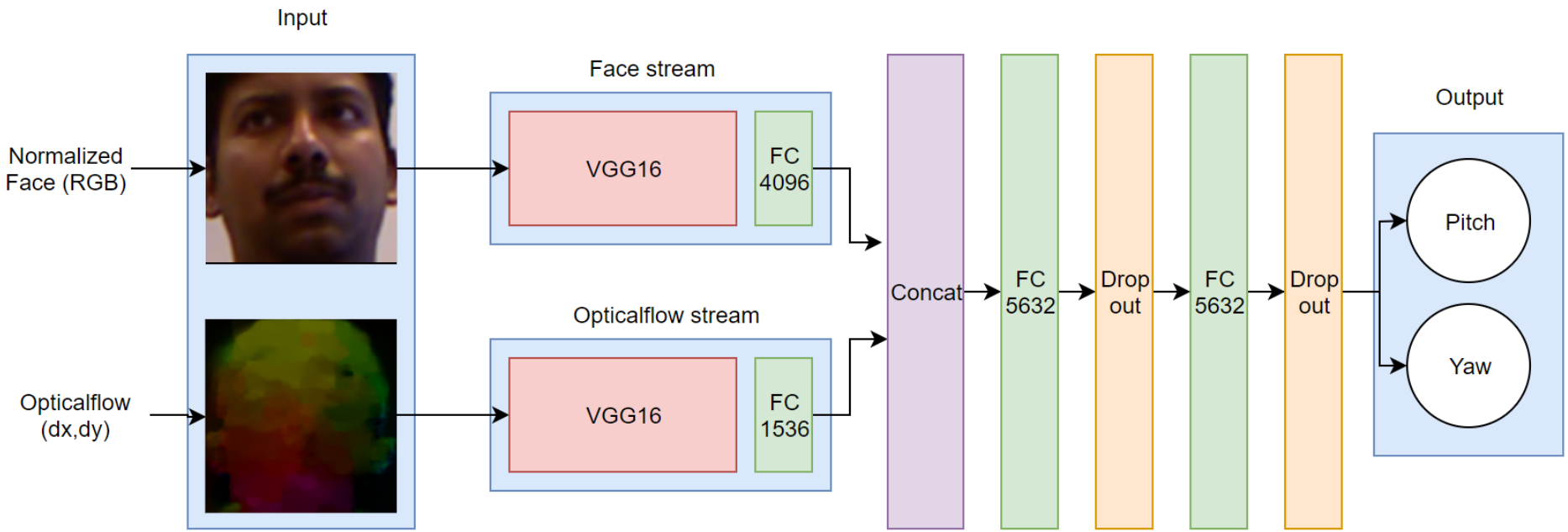



Figure: Architecture of the two stream network model for face and optical flow.

# Three streams

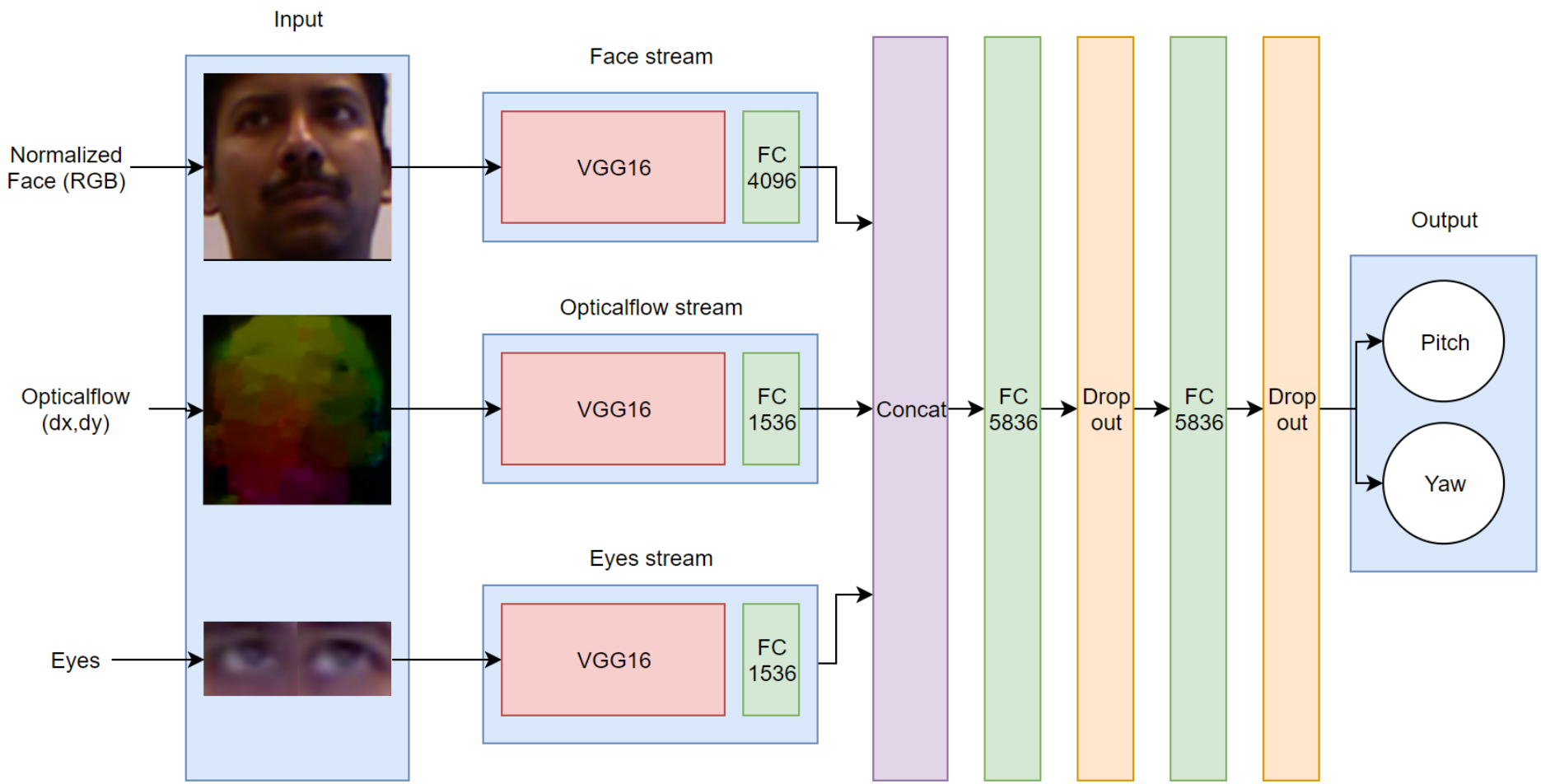



Figure: Architecture of the three stream network model for face, eyes and optical flow.

# Recurrent model

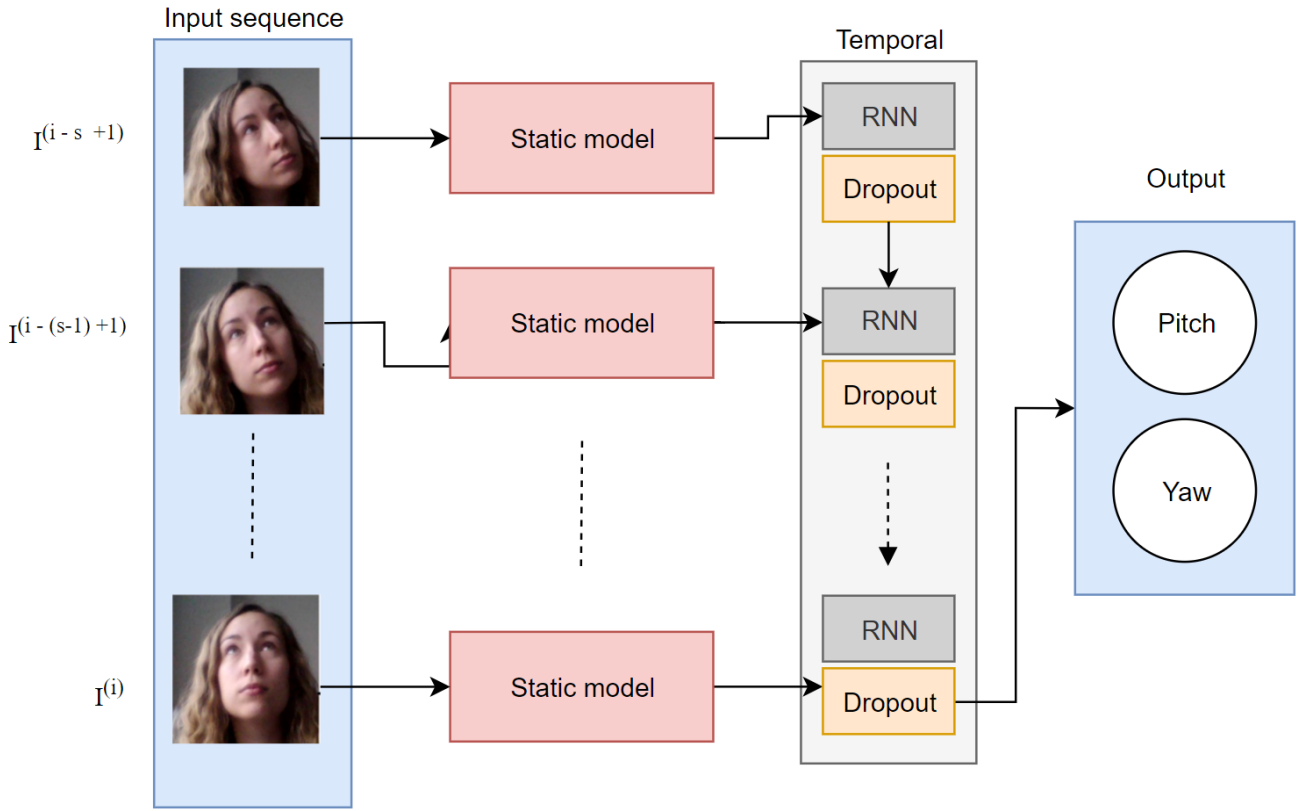



Figure: Recurrent model used in the temporal experiments. Face images extracted from [4].

# Outline

# Experimental protocol

- We have to define a **loss function**:

$$\text{angular error} = arccos\left(\frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}\right) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

- We did **four fold** cross validation for the 16 subjects of the EYEDIAP dataset.

# Baseline results

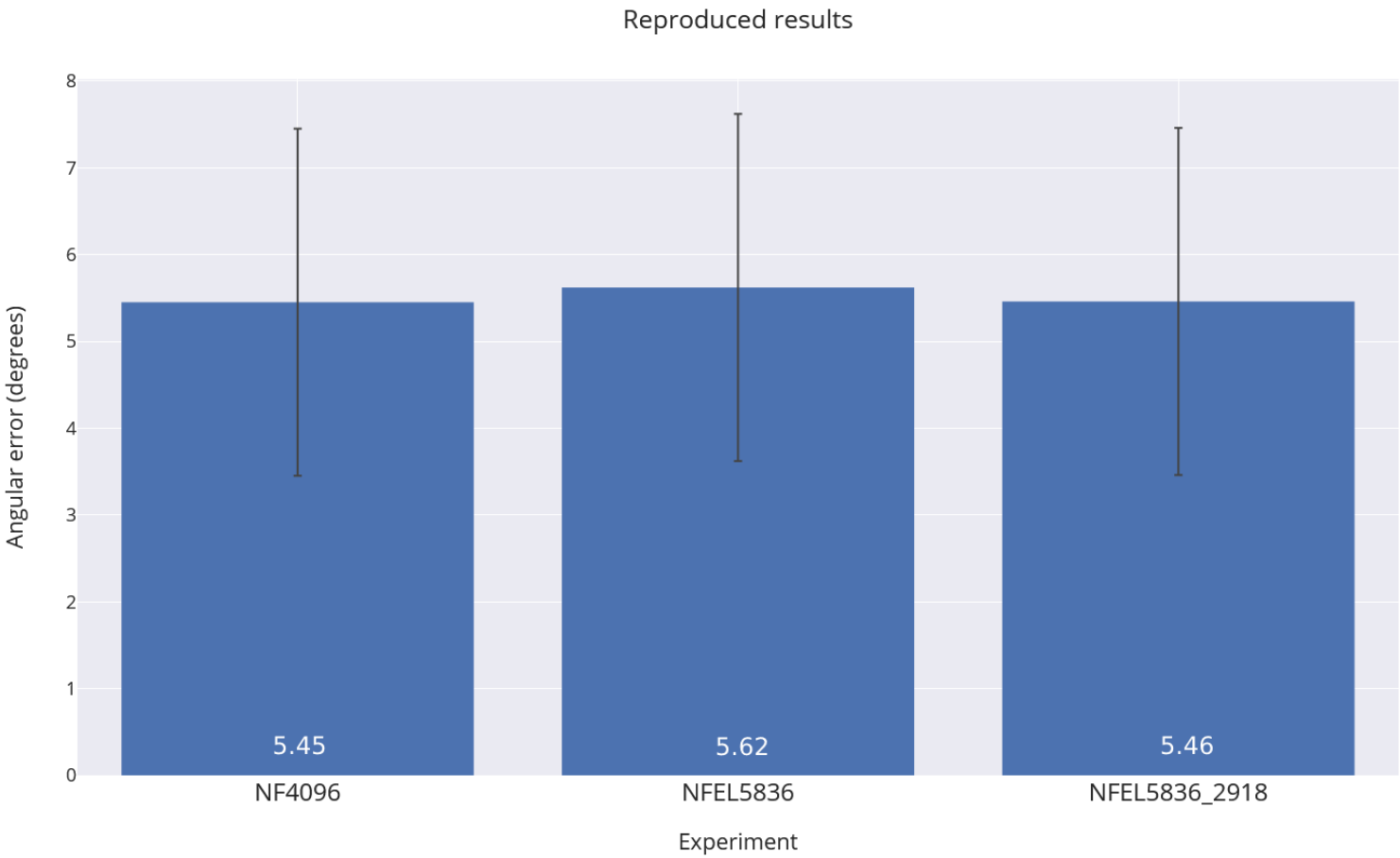



Figure: Means of the results of the folds of the experiments that we reproduced.

# Optical flow results

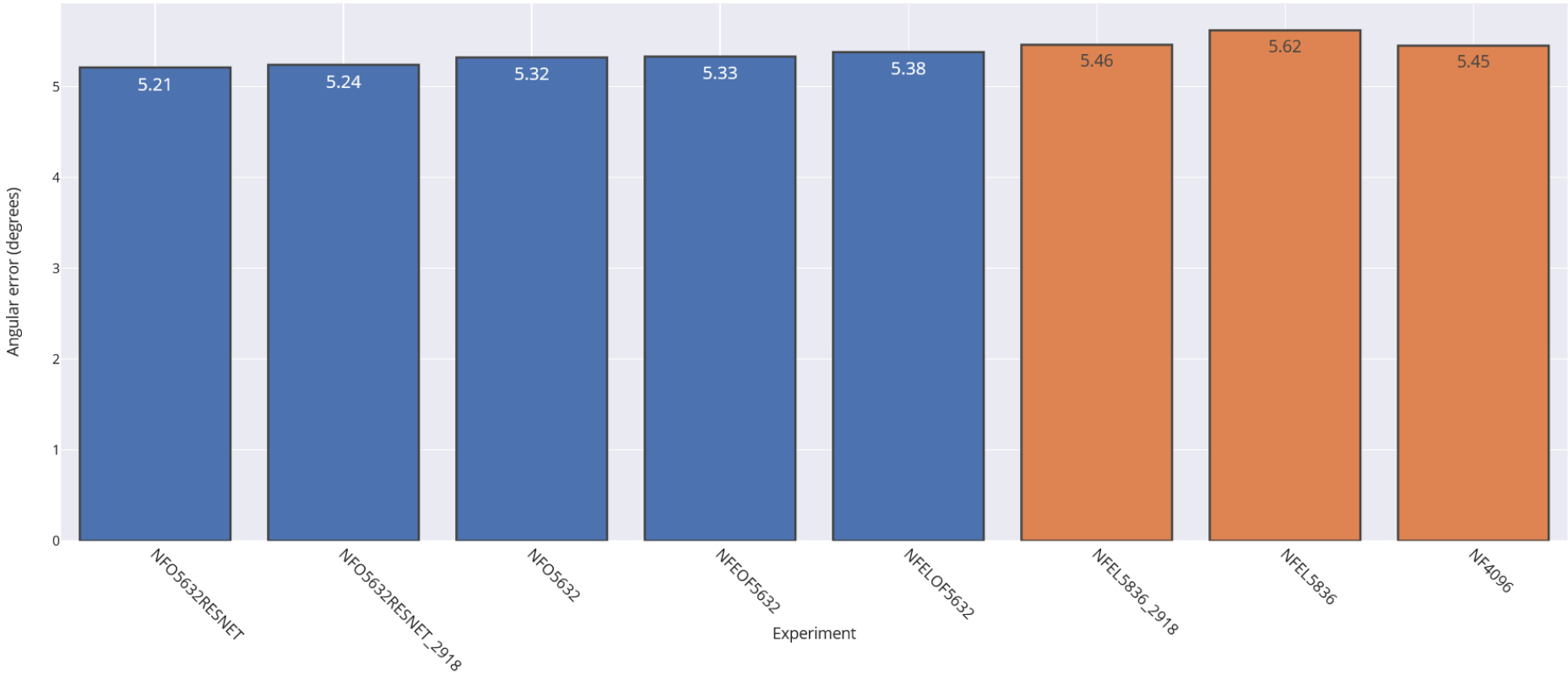



Figure: Mean angular error of all the frames in the dataset for each experiment.

# Recurrent experiments

We tried two models:

- A recurrent model with a GRU as a temporal layer that uses as base model the **NFEL5832_2918** (Normalized face, eyes, landmarks with 5836 neurons in the last fully connected layer).
  - Mean angular error of 5.29
- A recurrent model with a GRU as a temporal layer tat uses as base model the **NFO5632** (Normalized face, optical flow face, with 5632 neurons in the last fully connected layer).
  - Mean angular error of 5.28

# Outline

# Median filter

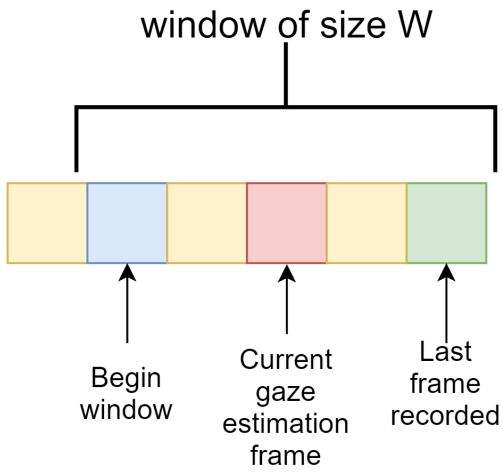



Figure: Illustrating the sliding window that we use in the filter.

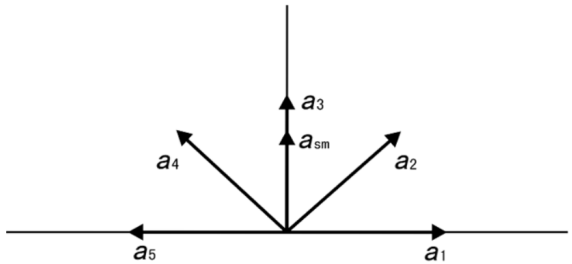



Figure: Comparison of the vector median vector and the scalar median vector. (from [1])

# Median filter Results

| Fold | Window size | | | |
|---|---|---|---|---|
| | Without filter | 3 | 5 | 7 |
| 1 | 4.246263 | 4.2088714 | 4.189559 | 4.190738 |
| 2 | 5.624934 | 5.597168 | 5.578781 | 5.586626 |
| 3 | 4.3602552 | 4.327826 | 4.3174233 | 4.3054256 |
| 4 | 6.570273 | 6.5285983 | 6.516427 | 6.5036163 |
| mean | 5.2 | 5.17 | 5.15 | 5.15 |

Table: Validation angular error when a median filter is applied to the experiment NFO5632RESNET in the four folds.

# Outline

# Eye Ablation studies preprocessing

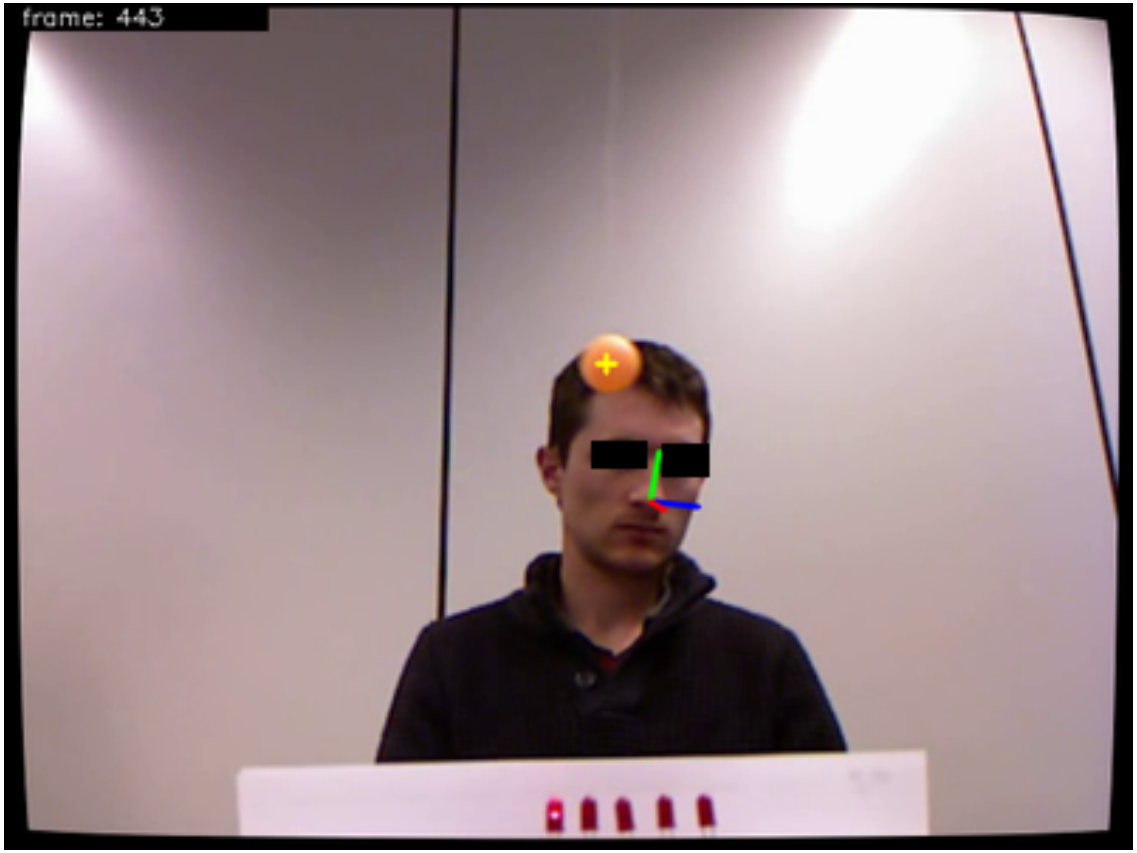



Figure: Subject 1_A_FT_M of the EYEDIAP dataset with eyes ablation.

# Eye Ablation results

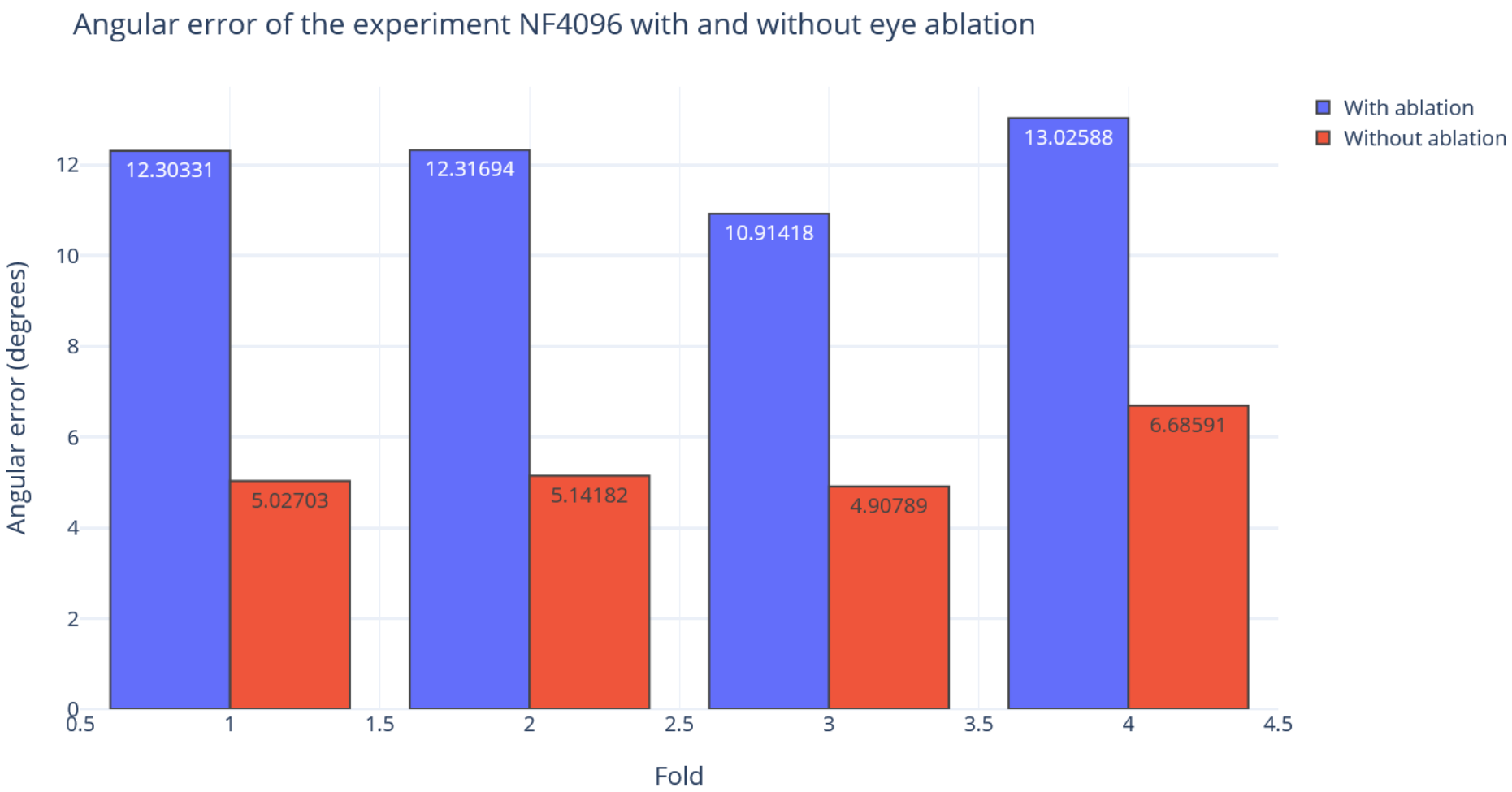



Figure: Comparison of the angular error for the experiment NF4096 between the original dataset and the dataset with eye ablation.

# Outline

# Ethics

- To obtain economical revenues
  - **Private companies** can do gaze analysis to determine **people interests** and adapt establishments to obtain **economic revenues**.

# Ethics

- To obtain economical revenues
  - **Private companies** can do gaze analysis to determine **people interests** and adapt establishments to obtain **economic revenues**.
  - Analyze how **two persons** are interacting and infer the **type of conversation** they are having.

# Ethics

- To obtain economical revenues
  - **Private companies** can do gaze analysis to determine **people interests** and adapt establishments to obtain **economic revenues**.
  - Analyze how **two persons** are interacting and infer the **type of conversation** they are having.
- To help people
  - Using gaze you could make applications that can detect if you are having a sad moment, or even **suicidal thoughts**.

# Ethics

- To obtain economical revenues
  - **Private companies** can do gaze analysis to determine **people interests** and adapt establishments to obtain **economic revenues**.
  - Analyze how **two persons** are interacting and infer the **type of conversation** they are having.
- To help people
  - Using gaze you could make applications that can detect if you are having a sad moment, or even **suicidal thoughts**.
  - It could help to improve your **communication skills** also.

# Ethics

- To obtain economical revenues
  - **Private companies** can do gaze analysis to determine **people interests** and adapt establishments to obtain **economic revenues**.
  - Analyze how **two persons** are interacting and infer the **type of conversation** they are having.
- To help people
  - Using gaze you could make applications that can detect if you are having a sad moment, or even **suicidal thoughts**.
  - It could help to improve your **communication skills** also.

Governs have to legislate about the advances made with artificial intelligence so that they are used mainly for beneficial purposes for the world and not misused

# Outline

# Conclusions and future work

- Conclusions
  - We saw that the addition of optical flow to the models improved the angular error until a 7%.
  - We analyze the effects of applying a **median filter** to the output vector using several window sizes.
  - An **Eye ablation study** has also been carried out.
- Future work
  - Best fine tunning of the **hyperparameters**.
  - Recurrent model with **ConvLSTMs**.
  - **3DCNN** to encode deep features.

# Master thesis

## Spatio-temporal Gaze Estimation for Human-Machine and Human-Human Interaction

Alberto López Sánchez
Director: Sergio Escalera
Codirector: Cristina Palmero

Master in artificial intelligence
Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat Rovira i Virgili

January 31, 2020

# Bibliografía I


Yike Liu. "Noise reduction by vector median filtering". In: *GEOPHYSICS* 78.3 (2013), pp. V79–V87. DOI: 10.1190/geo2012-0232.1. eprint: https://doi.org/10.1190/geo2012-0232.1. URL: https://doi.org/10.1190/geo2012-0232.1.

Kenneth Funes Mora, Florent Monay, and Jean-Marc Odobez. "EYEDIAP: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras". In: Mar. 2014, pp. 255–258. DOI: 10.1145/2578153.2578190.

Adrian Bulat and Georgios Tzimiropoulos. "How Far are We from Solving the 2D and 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks)". In: *2017 IEEE International Conference on Computer Vision (ICCV)* (Oct. 2017). DOI: 10.1109/iccv.2017.116. URL: http://dx.doi.org/10.1109/ICCV.2017.116.

# Bibliografía II

Cristina Palmero et al. "Recurrent CNN for 3D Gaze Estimation using Appearance and Shape Cues". In: *CoRR* abs/1805.03064 (2018). arXiv: 1805.03064. URL: http://arxiv.org/abs/1805.03064.